Camille Prime-Claverie
Université Paris Ouest Nanterre La Défense – Nanterre, France ; CRIS-SERIES.
Stéphanie Pouchot
Université de Lyon, France ; Université Claude Bernard Lyon 1, ELICO, EA 4147.

Archives ouvertes et banques de données commerciales : quelle visibilité pour la recherche en sciences de l'information et de la communication ?

Résumé : Grâce à une étude quantitative de corpus textuels (notices bibliographiques), la recherche présentée propose de confronter la visibilité de la recherche en sciences de l'information et de la communication émanant d'une archive ouverte (@rchiveSIC) et d'une base de données commerciale (Francis).

Abstract: Based on a quantitative study of corpus (bibliographic records), this research confronts research visibility in information studies and communications disseminated through an open archives (@rchiveSIC) and a commercial database (Francis).

1. Problématique et contexte général

La révolution numérique de ces vingt dernières années a induit de profondes modifications dans l'accès à la littérature scientifique comme dans les pratiques de publication des chercheurs. Les systèmes d'autoarchivage en ligne permettent en effet aux chercheurs de mettre à disposition leurs publications sur le Web sans passer par les canaux habituels de l'édition scientifique.

En France, la volonté de soutenir les initiatives d'archives ouvertes s'affirme dès le début des années 2000 avec la création du Centre pour la Communication Scientifique Directe (CCSD) du Centre National de la Recherche Scientifique (CNRS). Depuis la déclaration de Budapest en 2002 et l'appel international de Berlin lancé en 2003 pour la constitution et la diffusion d'archives ouvertes, le mouvement OAI (Open Archive Initiative) s'accélère : le nombre de sites d'archives ouvertes a triplé entre 2008 et 2009 (Schöpfel, 2009).

Les pratiques des chercheurs restent néanmoins disparates. Si les astrophysiciens, avec la fameuse base ArXiv, ont très tôt compris l'intérêt des archives ouvertes et se sont rapidement et massivement approprié cet outil, cela est moins évident dans les champs des sciences humaines et sociales, tant du côté des publiants que de celui des lecteurs (Bosc, 2005). Il paraît alors légitime de s'interroger sur la représentativité de l'activité scientifique relative à un champ de recherche, d'une part dans les archives ouvertes et d'autre part dans les banques de données commerciales. La présence des articles dans ces deux types de bases dépendant des pratiques des chercheurs, comment envisager le dynamisme et appréhender les thématiques de recherche émergentes d'un domaine d'études donné ?

2. Objectifs de la communication

Ce travail de recherche s'inscrit dans le cadre du projet PIC (Plateforme Intégrée de Consultation et manipulation de corpus textuels pour les SHS) et est soutenu par le TGE ADONIS[i] (Très Grand Equipement Accès unifié aux DOnnées et documents NumérIques

des Sciences humaines et sociales). Il s'agit d'une infrastructure dédiée aux problématiques relatives au numérique (accès, publication, utilisation, conservation) pour les sciences humaines et sociales (SHS).

Dans cette communication, nous présentons les choix méthodologiques et les résultats d'une partie du projet PIC centrée sur l'étude de l'état de la recherche dans le domaine des sciences de l'information et de la communication (SIC). Nous nous sommes intéressées à la représentation pouvant en être construite à partir d'une banque de données commerciale et d'une archive ouverte. Cette contribution s'inscrit donc dans l'axe 3 du congrès, « L'organisation de l'information ».

Notre objectif de recherche est ici d'envisager si, devant l'essor des nouvelles formes de publications en accès libre, les banques traditionnelles donnent toujours une représentation thématique satisfaisante des activités de recherche en SIC ou si, *a contrario*, il devient incontournable de prendre en compte les archives ouvertes pour avoir une vision réaliste de la recherche en SIC. Il s'agit notamment de déterminer s'il existe une superposition des thématiques de recherche ressortant de ces deux canaux de diffusion ou si certains des axes de recherche, en particulier les axes émergeants, sont mieux représentés d'un côté ou de l'autre.

3. Cadre théorique et méthodologique

Nous avons adopté une démarche quantitative pour cette étude et travaillé sur un ensemble de données textuelles (notices bibliographiques) extraites d'@rchiveSIC[ii] et de Francis[iii].

Délimitation du périmètre de l'étude

Afin de pouvoir comparer les différents thèmes émanant de bases de données, il est important de disposer, pour chaque document, d'informations provenant de l'indexation, par exemple, des indices de classement, des descripteurs, *etc*. Comme d'autres auteurs, notamment (Roche, 07), nous soulignons la pauvreté des plans de classement utilisés pour l'indexation, pour la plupart des banques de données et des archives ouvertes. Les analyses portant sur de hauts niveaux d'agrégation, c'est-à-dire sur un découpage disciplinaire général, ne permettent pas d'identifier finement des axes de recherche spécifiques. C'est pourquoi notre choix s'est porté sur @rchiveSIC plutôt que sur HAL-SHS, @rchiveSIC présentant un plan de classement thématique plus détaillé. Concernant les bases commerciales, nous avons opté pour Francis qui utilise un vocabulaire contrôlé pour l'indexation ainsi qu'un plan de classement.

Constitution et organisation du corpus

En avril 2009, le moissonnage de la base @rchiveSIC par le protocole OAI-PMH a permis de constituer un ensemble de 892 notices selon le standard du Dublin Core. @rchiveSIC ayant vu le jour en mai 2002, la majorité des articles déposés ont été publiés entre 2000 et 2009. Nous avons par conséquent choisi d'interroger Francis sur cette même période et limité l'interrogation aux catégories « Sciences de l'information » et « Sociologie de la communication et des mass media / Linguistique » avec au moins un des auteurs affilié en France. Nous avons ainsi colligé 3953 notices.

Pour permettre la caractérisation thématique de nos deux sous-corpus, une partie du

travail a consisté à organiser les données au sein d'une base SQL.

Méthodologie et contribution de la recherche

Dans les études scientométriques, la caractérisation thématique des corpus se fait en général par l'affectation de sujets aux articles en fonction des revues dans lesquelles ils sont publiés et non en fonction du contenu même de l'article. Cela peut entraîner un certain nombre de confusions et imprécisions. Notre objectif ici est de déterminer dans quelle mesure un autre type de caractérisation est possible. Nous proposons d'adopter le point de vue des fournisseurs des bases retenues, c'est-à-dire d'utiliser comme point d'entrée les plans de classement d'@rchiveSIC et Francis.

Pour les deux sources, nous avons procédé à un double décompte des publications pour chaque catégorie : d'une part, un compte de présence, d'autre part, un compte fractionnaire. Pour le compte de présence, un article présent dans X catégories compte pour 1 dans chaque catégorie. Le compte fractionnaire est pondéré, un article présent dans X catégories comptant pour 1/X dans chacune des catégories concernées.

En outre, les plans de classement d'@rchiveSIC et Francis étant différents en termes de catégories proposées et de profondeur, la comparaison demeure délicate. Il est alors nécessaire d'aligner les catégories c'est-à-dire de rechercher des correspondances entre les deux plans de classement. Par exemple la classe « Aspects juridiques : propriété intellectuelle, responsabilité du producteur. Ethique » s'approche de la catégorie « Droit de l'information/communication » présente dans @rchiveSIC.

L'alignement des classes reste cependant une opération difficile, certaines catégories d'une des sources se retrouvant morcelées dans plusieurs catégories de l'autre, ce qui particulièrement vrai pour la partie documentation et sciences de l'information. De même, si @rchiveSIC sépare muséologie d'une part et cinéma, art, esthétique de l'autre, Francis regroupe ces thématiques.

Nous avons néanmoins pu effectuer un certain nombre de calculs et comparé les résultats issus des deux sous-corpus. Si certaines catégories sont représentées de manière équivalente dans les deux bases étudiées (par exemple la bilbiométrie-scientométrie qui compte un peu plus de 2% des articles dans les deux bases), il existe des différences marquées. Le tableau ci-dessous met en évidence, pour quelques unes des catégories, la répartition des publications en compte fractionnaire selon les deux plans de classements.

Catégorie @rchiveSIC	**% articles (compte fractionnaire)**	**Catégorie Francis**	**% articles (compte fractionnaire)**
Sociologie de l'information/communication	8,87%	Sociologie de l'information et de la communication	0,09%
Communication et information scientifique	8,69%	Communication et information scientifique	0,03%
Education, formation	4,86%	Education, formation	0,03%
Médias de masse	4,12%	Sociologie de la communication et des mass media. Linguistique	22,67%
Economie, industries culturelles	2,36%	Economie, industries culturelles	0,13%

Droit de l'information/communication	1,20%	Aspects juridiques : propriété intellectuelle, responsabilité du producteur. Ethique	7,30%

Tableau 1 – Compte fractionnaire des publications pour quelques catégories @rchiveSIC et Francis

Un des résultats de cette étude est que les thématiques relevant de la communication et de l'information scientifique, de l'économie et des industries culturelles ou encore de l'éducation et de la formation sont plus présentes dans @rchiveSIC que dans Francis, qui favorise davantage la visibilité de la production en documentation et sciences de l'information. Nous pouvons noter par ailleurs la moindre représentativité des aspects juridiques dans @rchiveSIC.

La globalité des résultats statistiques concernant la répartition des articles permet de tirer des conclusions sur la représentativité des SIC à travers une archive ouverte et une base de données commerciale. Ils induisent également un certain nombre d'interrogations et permettent d'envisager de nouvelles pistes de réflexion.

4. Perspectives

A quoi sont dus ces recoupements, ces différences ? Cela relève-t-il des pratiques de publication et d'archivage des chercheurs ? De leurs pratiques informationnelles ? Ou les vues sont-elles simplement les conséquences des différents plans de classement ? Francis adoptant un découpage disciplinaire classique, l'aspect transdisciplinaire des sciences de l'information et de la communication est peu pris en compte. Ainsi, une partie des publications de la discipline peut se retrouver à différent niveaux de la classification et aurait pu nous échapper.

Ces travaux contribuent à mieux cerner les pratiques de publication et d'auto-archivage des chercheurs en SIC. Par ailleurs, l'approche proposée et la méthodologie sont généralisables : elles permettront d'étudier d'autres champs disciplinaires à partir de données issues d'autres bases.

De manière plus large, cette recherche questionne les techniques d'évaluation des structures de recherche s'appuyant sur des indicateurs quantitatifs basés sur des banques de données payantes et contribue au débat sur les indicateurs de la recherche.

5. Quelques références bibliographiques

Bar-Ilan, Judit. 2008. Informetrics at the beginning of the 21st century – A review. *Journal of Informetrics*, Volume 2, Issue 1, Pages 1-52.

Bosc, Hélène. 2005. Archives ouvertes : quinze ans d'histoire. *Les archives ouvertes : enjeux et pratiques. Guide à l'usage des professionnels de l'information.* C. Aubry, J. Janik (eds.), Paris : ADBS, Pages 27-54.

Roche, Ivana, Claire François et Dominique Besagni. 2007. Détection de techniques prometteuses à partir de méthodes bibliométriques, *CIDE* [En ligne], *Session Bibliométrie, CIDE 10*, mis à jour le : 10/09/2008, URL : http://172.16.128.67:50010/cide/index.php?id=229.

Schöpfel, Joachim et Hélène Prost. 2009. Les statistiques d'utilisation d'archives

ouvertes. Etat de l'art. *Ressources électroniques académiques : mesures et usages, Colloque international*, 26-27 Novembre 2009, Lille.

[i] http://www.tge-adonis.fr/ (site consulté le 29/03/10).
[ii] Archive ouverte en sciences de l'information et de la communication, http://archivesic.ccsd.cnrs.fr/ (site consulté le 29/03/10)
[iii] Base de données bibliographiques en sciences humaines et sociales produite par l'INIST, voir description sur http://ingenierie.inist.fr/spip.php?article1 (site consulté le 29/03/10)